View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Munin - Open Research Archive

Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos

Konstantin Pogorelov[1,2], Olga Ostroukhova[7], Mattis Jeppsson[6], Håvard Espeland[6], Carsten Griwodz[1,2], Thomas de Lange[1,5], Dag Johansen[3], Michael Riegler[1,2,4] and Pål Halvorsen[1,2,4]

[1]*University of Oslo, Norway* [2]*Simula Research Laboratory, Norway* [3]*UiT - Artic University of Norway*
[4]*Simula Metropolitan Center for Digital Engineering, Norway*
[5]*Oslo University Hospital, Norway* [6]*ForzaSys AS, Norway*
[7]*Research Institute of Multiprocessor Computation Systems n.a. A.V. Kalyaev, Russia*

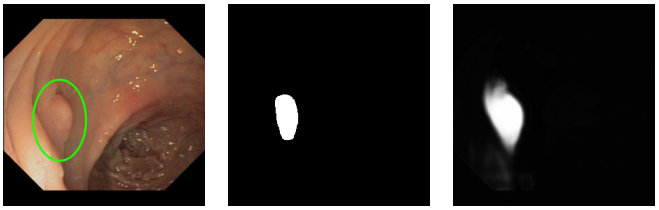
***Abstract*—Video analysis including classification, segmentation or tagging is one of the most challenging but also interesting topics multimedia research currently try to tackle. This is often related to videos from surveillance cameras or social media. In the last years, also medical institutions produce more and more video and image content. Some areas of medical image analysis, like radiology or brain scans, are well covered, but there is a much broader potential of medical multimedia content analysis. For example, in colonoscopy, 20% of polyps are missed or incompletely removed on average [1]. Thus, automatic detection to support medical experts can be useful. In this paper, we present and evaluate several machine learning-based approaches for real-time polyp detection for live colonoscopy. We propose pixel-wise localization and frame-wise detection methods which include both handcrafted and deep learning based approaches. The experimental results demonstrate the capability of analyzing multimedia content in real clinical settings, the possible improvements in the work flow and the potential improved detection rates for medical experts.**

***Index Terms*—medical video analysis, machine learning, deep learning, image features, performance**

I. Introduction

Hospitals record and collect a huge amount of multimedia data which needs to be stored and analyzed, both on-the-fly and offline. One example is gastrointestinal (GI) tract examinations where large numbers of videos are collected, i.e., by an endoscope controlled by a medical expert. Making the future GI examinations more efficient and cost-effective is also a huge societal challenge as about 2.8 millions of new esophagus, stomach and colorectal cancers are detected yearly in the world with a mortality of about 65% [2]. All have a significant impact on the patients' health-related quality of life. Consequently, gastroenterology is one of the most significant medical branches. Colorectal cancer is the third most common cause of cancer mortality for both women and men, and it is a condition where early detection is important for survival. For example, a patient is going from a low 10-30% 5-year survival probability if detected in later stages of the disease to a high 90% survival probability in early stages [3].

Colonoscopy is considered to be the gold standard for the examination of the colon for early detection of cancer and precancerous pathology. However, it is not an ideal screening test. Polyps, which are abnormal growth of tissue projecting from a mucous membrane (see Figure 1(a)), are often predecessors of colorectal cancers, and are therefore important to detect early. However, on average, 20% of polyps are missed or incompletely removed, i.e., the risk of getting cancer largely depend on the endoscopists ability to detect polyps [1]. It is also a demanding procedure requiring a significant time investment from the medical professional, and the procedure is unpleasant and can cause great discomfort for the patient [4]. Furthermore, there are high costs related to the procedure. Norway has an average cost of about $450 per examination. In the US, colonoscopy is the most expensive cancer screening process with an average of $1,100 per examination, i.e., an annual cost of $10 billion dollars [5].



(a) Input frame (b) Ground truth mask (c) Segmentation mask

Fig. 1. Example of a polyp marked with a green circle (a), a corresponding polyp localization ground truth mask (b) and our output segmentation mask using GAN (c). Images taken from the CVC-968 [6].

In the area of image analysis and object detection, machine learning, and especially deep learning, has been very popular, also in the field of medicine, in the recent years. Deep learning algorithms are based on neural networks that use recently developed training techniques to train their models. They are basically an abstracted representation of data points. The representation is made on a high-level, and multiple layers for processing the networks are used to reach higher complexity. The different layers can learn different abstraction levels of the data using input of previous layers until they reach a final layer, which makes the final decision for the class. The new training techniques for deep learning were mainly made possible because of the emergence of GPU computing, which enables training of the networks in a reasonable amount of time. On the other side, the disadvantages include a very long training time, classification boundaries are hard to explain

(why one data point is put in this class), and they are very data driven [7], [8].

Automatic detection of polyps is in general well researched, and there are many publications on the topic. Related work indicates with a sensitivity and specificity close to one that the problem is solved. Nevertheless, there are still several open challenges, e.g., the evaluation of existing approaches is often performed on small and non-publicly available datasets. Medical datasets also have the challenge that they usually contain many true negative examples, but not so many true positives. Furthermore, a very important open question is how generalizable the proposed methods are. Generalization is a vital ability of a model trained on a dataset from one hospital to be applied in another hospital, e.g., using a different type of equipment (endoscope). Therefore, in this paper, we are addressing the challenges arising due to limited datasets and generalizability of models which both are common problems in medical multimedia scenarios [9].

The main contributions of this paper are proposing and testing different approaches to overcome the problems concerning generalization of models and limited datasets in terms of size and sample equality, and to propose a best approach for detection and localization of findings for medical image analysis. Our best working approach outperforms by far all our own and other tested approaches, and does at the same time not need a large amount of training data. Furthermore, it achieves good performance across datasets and does not need negative examples for training. With respect to dataset size and generalizability, we conclude that one proposed detection and localization model can be used across different datasets and different equipment and it is able to perform efficiently using very low amount of training samples. With our best working approach based on a generative adversarial network (GAN), we reach a detection specificity of 94% and an accuracy of 90.9% with only 356 training [6] and 6000 test [10] samples captured by different equipment.

The rest of the paper is organized as follows: first, we give an overview of the related work in the field. This is followed by a description of our methods, which we next experimentally evaluate. Finally, we conclude the paper and give directions for future work.

II. Related work

Recently, frame-wise detection and in-frame localization of colon polyps have been picked up as a research topic by many scientists in the medical imaging area, but lately also in the multimedia community. Approaches in context with automatic detection or localization of polyps in videos taken from colonoscopies can be divided into hand-crafted feature based, re-training or fine-tuning of existing and trained from scratch deep learning architectures.

In hand-crafted feature based approaches for detection, researchers extract features such as global or local image features (texture, edge or color based) from the frames and use them within different machine learning algorithms such as random forest (RF) or support vector machines (SVM) [11], [12]. The best working hand-crafted detection approaches are [13] and [9] with both precision and recall above 90%. The first approach [13] relies on edge and texture features whereas the latter [9] uses several different global image features. For localization, the best working approaches from Yuan et. al. [14], who use a bottom-up and top-down saliency approach, and from Wang et. al. [13], where they use edge and texture features. Usually, localization approaches can also be used for frame-wise detection.

Reusing already existing deep learning architectures and pre-trained models leads to very good results in for example the Imagenet classification tasks. Retraining architectures from scratch in the context of colonoscopies leads to reasonable good results, but the limited size of medical datasets is a problem for these approaches. For pre-trained models, even if their categories are quite different compared to the medical use case, it has been shown that they can be used in the context of polyp detection and localization tasks [15], [16], and that they often outperform hand-crafted approaches [17], [18].

In [19], a 3D convolutional neural network (CNN) architecture approach is presented for polyp detection. The method is also compared to hand-crafted and 2D CNN approaches, and it is shown that different approaches perform well for different sub-tasks. For example, the hand-crafted feature approach is working well for true negative detection. The best performance is reached with a fusion of all investigated approaches. Moreover, Pogorelov et. al. [20] and Riegler et. al. [21] compare different localization approaches (hand-crafted and deep learning). The conclusion is that pre-trained and fine-tuned deep learning models outperform other approaches, but that they are far away from being ready for clinical use (usually a sensitivity and specificity above 85% is considered as the borderline [22]).

In general, recent related work reports very promising results in terms of evaluation metrics, i.e., both recall (also called sensitivity) and specificity close to one. Nevertheless, most of the approaches are tested on small and non-publicly available datasets. Furthermore, the problem of medical datasets is that they usually contain many negative examples, but not so many positives is not well researched. Another open question is how generalizable the proposed methods are, meaning can a model trained on a dataset from one hospital be applied in another hospital. These are questions that we are addressing in this paper.

III. Methodology

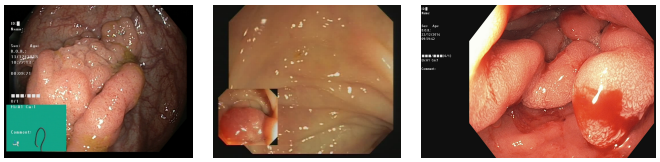
A. Pixel-wise segmentation/localization approach

The first presented segmentation approach is able to pixel accurate mark the polyp in the given frame. We use generative adversarial networks (GANs) to perform the segmentation. GANs [23] are machine learning algorithms that are usually used in unsupervised learning and are implemented by using two neural networks competing with each other in a zero-sum game. We used a GAN model architecture initially developed for the retinal vessel segmentation in fundoscopic images, called V-GAN, as basis for our polyp segmentation approach.

The V-GAN architecture [24] is designed for RGB images and provides a per-pixel image segmentation as output. To be able to use the V-GAN architecture in our polyp segmentation approach, we added an additional output layer to the generator network that implements an activation layer with a step function which is required to generate the binary segmentation output. Furthermore, we added support for gray-scale and RGB color space data shapes for the input layers of the generator and discriminator networks including an additional color space conversion step. Gray-scale support was added to be able to use a single value per pixel input in order to reduce the network architecture complexity and to speed up the model training and data processing parts.

Data preparation. The frames used in this research is obtained from the standard endoscopic equipment and can contain some additional information fields related to the endoscopic procedure. Some types of the fields (see Figure 2), integrated into resulting frames showed to the doctor and captured by the recording system, can confuse detection and localization approaches, and it leads to frame miss-classification (green navigation box) or false positive detection (captured frame with polyp). We have implemented a simple frame preparation procedure that consists of three independent steps: a black border removal (including patient-related text fields), a green navigation localizer map masking and a captured still frame masking. All the removed and masked regions are excluded from further frame analysis.

Data augmentation. Due to a limited number of frames with the detailed ground truth masks, we implemented a data augmentation scheme used in the training process of the GAN. For the experiments presented here, we used only rotation and flipping of frames. Rotation was performed independently with 20° steps for the original. Together with the in-horizontal-direction-flipped frames, we added 35 new frames complementary to the original ones.



(a) Navigation (b) Captured frame (c) Patient information

Fig. 2. Examples of the different auxiliary information fields integrated into recorded frame: a colonoscope navigation localizer (a), a captured still frame (b) and a patient-related information (c). Images taken from CVC-968 [6] and Kvasir [10].

B. Frame-wise detection approaches

Frame-wise detection approaches are designed to detect the target object on a per-frame level, i.e., in our GI scenario, detect if there is a polyp in the frame or not. For frame-wise detection, we propose different methods. We conducted experiments using various configurations of our main methods. The main methods are hand-crafted global features (GF-D), re-training and fine-tuning on existing deep learning architectures (RT-D) and a variation of the GAN approach (GAN-D) that was also used for the pixel-wise segmentation.

TABLE I
ARCHITECTURES AND CONFIGURATIONS USED FOR RT-D. WE HAVE USED THE *rmsprop* AND *SGD* OPTIMIZERS IN STEPS 1 AND 2, RESPECTIVELY, 50 EPOCHS AND A BATCH SIZE OF 32.

Method	Architecture	Step 2: frozen from layer	Image size
RT-D-Xcept	Xception [29]	26	299x299
RT-D-VGG19	VGG19 [30]	5	224x224
RT-D-ResNe	ResNet50 [31]	50	224x224

GF-D. For the GF method, we extracted handcrafted global features (describing the image on a global level, e.g., texture, color distribution, etc.) using the LIRE framework [25]. The features that we used are Joint Composite Descriptor, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and Pyramid Histogram of Oriented Gradients. We performed early fusion by combining all extracted features resulting in a feature vector with the size of 1186. For the classification, we used a Logistic Model Tree (LMT) classifier from the Weka machine learning library [26].

RT-D. For the RT method, we implemented a re-training and fine-tuning approach and used it with three well known and working architectures. For all architectures, we used models trained on the Imagenet dataset for starting weights. The approach for RT-D works in two steps. First, we freeze all layers of the architecture and train only the base layers. After that, we unfreeze certain layers and fine-tune the network. Which blocks are un-freezed for the second step is decided via a Bayesian optimization algorithm [27] which runs for 20 iterations. To find good working optimizers, number of epochs and batch sizes for the different architectures, we also used Bayesian optimization for 20 iterations including all architectures. This lead to values that gave good overall results and could be used for all architectures to achieve better comparability. Details about the exact configurations and architectures used can be found in Table I. The dataset used for the optimization step is public available and details can be found in [28].

GAN. The GAN detection approach utilizes a simple threshold activation function, which takes the number of positively marked pixels in the frame as input. In the validation experiments performed using different datasets, we evaluated the activation thresholds from one pixel to a quarter of the frame. The best detection results were achieved with a threshold value of 50 pixels, which has been used for the detection experiments for the development and test set and confirms high performance of the GAN-based localization approach.

C. Block-wise segmentation/localization approach

The second localization approach is our attempt to utilize frame-wise detection algorithm for localization purposes. We have applied the RT-D method to the set of sub-frames generated from the training and test sets. Sub-frames (blocks) are generated using sliding square window with 66% overlap with the neighbor sub-frames. We have tested different window

TABLE II
OVERVIEW OF THE DATASETS USED IN THE EXPERIMENTS.

Dataset	Training	Test	# Frames	# Polyp frames	# Normal frames
CVC-356	X	X	1,706	356	1,350
CVC-612	X	X	1.962	612	1,350
CVC-968	X	X	2.318	968	1,350
CVC-12k	-	X	11,954	10,025	1,929
Kvasir	-	X	6,000	1,000	5,000
Nerthus	X	-	1,350	-	1,350

sizes from 64x64 to 128x128 pixels. The best results were obtained using 128x128 windows size. The generated sub-frames are fed into the RT-D detection algorithm, and then, the processed sub-frames are grouped back into the frame. This results in a coarse localization map which is then used for the frame-wise detection. The detection is achieved by applying a simple threshold activation function, and we evaluated the activation thresholds ranging from 1 block to 50% of the frame blocks. The best detection results were achieved with a threshold value of 2 blocks.

IV. EXPERIMENTS

For the experiments, we use combinations of six different, publicly available datasets, namely CVC-356 [6], CVC-612 [32], CVC-968, CVC-12k [6], Kvasir [10] and parts of Nerthus [28] (see Table II for an overview). The CVC-356 and CVC-612 consist of 356 and 612 video frames, respectively. CVC-968 is a combination of CVC-356 and CVC-612. Each frame that contains a polyp comes with pixel-wise annotations in the CVC-356 and CVC-612 datasets. They are used for both training and testing in the localization performance evaluation experiments, and for the training only in the detection experiments. For the frame-wise detection approaches, except for the GAN-based approach, we also added the 1,350 class three frames with normal mucosa from the Nerthus dataset since normal mucosa examples for the negative class are required for our detection algorithms. The CVC-12k dataset contains 11,954 video frames. From these 11,954 frames, 10,025 contain a polyp and 1,929 show only normal mucosa. The polyps are not annotated pixel-wise, but with an oval shape covering the whole polyp (approximated annotation). For the Kvasir dataset, we included all classes except for the dyed classes (in a real world scenario something dyed is already detected by the doctor) leading to a dataset containing 1,000 frames with polyps, 5,000 without and only frame-wise annotations. The CVC-12k dataset is used as test set for block- and frame-wise detection and the Kvasir dataset for frame-wise detection.

A. Evaluation Metrics

For the evaluation of the experiments, we used the metrics precision (PREC), recall/sensitivity (SENS), specificity (SPEC), accuracy (ACC), F1 score (F1) and Matthew correlation coefficient (MCC). A detailed description and reasoning for the used metrics can be found in [10]. The localization metrics are calculated pixel- and block-wise using the provided binary masks of the ground truth.

TABLE III
VALIDATION RESULTS OF THE IN-FRAME PIXEL-WISE POLYP AREAS SEGMENTATION (LOCALIZATION) APPROACH EVALUATED USING DIFFERENT COMBINATIONS OF THE CVC-356 AND CVC-612 SETS FOR TRAINING AND TESTING.

Test set	Run	Train set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-612	LOC-356	CVC-356	0.819	0.619	0.984	0.946	0.706	0.684
CVC-356	LOC-612	CVC-612	0.723	0.735	0.981	0.965	0.729	0.710

B. Results

Table III depicts the performance evaluation results for the GAN-based pixel-wise segmentation approach. The best performance is achieved using the CVC-612 dataset for the training, which means, more training data improves the final results. An interesting observation is that the precision is higher with CVC-356 as training data. This might be an indicator that more training data makes the model more general, but less accurate. All in all, the validation using different datasets indicates that the approach works well, and the proposed localization algorithm can perform efficiently even with a low number of training samples available. This is important for our medical use-case scenario with a high diversity of objects and a limited amount of annotated data available. The initial localization experiments demonstrated more than 50% increase in performance of the localization using augmented training data, thus we have used augmented training data in all the pixel-wise localization experiments. A possible positive effect of test data augmentation with the following aggregation of the localization results will be subject of future research.

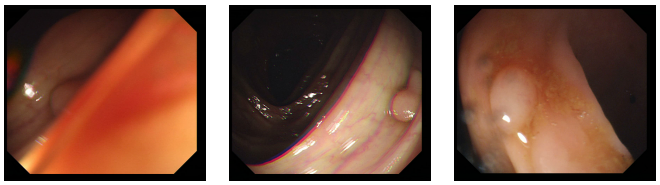
The results for the block-wise location approaches are presented in Table IV. The performance results obtained are especially interesting since all the approaches presented are trained with small amounts of training data without any negative examples (no normal mucosa frames at all). Furthermore, the CVC-12K dataset is heavily imbalanced which also makes it harder to achieve good results. For block-wise location via detection, the LOC-Xcept approach performs best for all the different training set sizes. It also indicates that a larger training dataset can lead to better results. The results for the LOC-ResNe approach confirm this with significant improvements when the training dataset size is increased. This is something that should be investigated in the future. Furthermore, the algorithm used to combine the results on the different sub-frames into one can be improved by, for example, using another machine learning algorithm to learn the best combinations.

The frame-wise detection results can be found in Table V. All approaches are trained on CVC-356, CVC-612 and CVC-968 training datasets and tested on the CVC-12k and Kvasir datasets. All in all, the GAN approach performs best on both datasets and within all variations of training datasets. The performance on the Kvasir dataset is better than on the

TABLE IV
PERFORMANCE OF THE BLOCK-WISE LOCALIZATION VIA DETECTION APPROACHES REPORTED PER METHOD AND USED TRAINING DATA.

Test set	Run	Training set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-12k	LOC-Xcept-356	CVC-356	0.475	0.203	0.966	0.868	0.285	0.250
	LOC-Xcept-612	CVC-612	0.528	0.289	0.961	0.874	0.374	0.328
	LOC-Xcept-968	CVC-968	0.584	0.257	0.972	0.880	0.357	0.333
	LOC-VGG19-356	CVC-356	0.257	0.292	0.874	0.799	0.273	0.158
	LOC-VGG19-612	CVC-612	0.266	0.489	0.799	0.759	0.344	0.228
	LOC-VGG19-968	CVC-968	0.232	0.406	0.800	0.750	0.295	0.166
	LOC-ResNe-356	CVC-356	0.723	0.003	0.999	0.871	0.006	0.044
	LOC-ResNe-612	CVC-612	0.469	0.054	0.990	0.869	0.098	0.125
	LOC-ResNe-968	CVC-968	0.536	0.248	0.968	0.875	0.340	0.306

CVC-12k dataset which is surprising since the Kvasir data is completely different from the CVC training data. Moreover, frames in the Kvasir dataset are captured using different and various hardware. This is a strong indicator that the approach is able to create a general model that is not just working well on the given data and that the CVC-12k dataset is very challenging. Some of the difficulties we could observe are for example screens in screens that show different parts of the colon, out of focus, frame blur, contamination, etc. (see for example Figures 2 and 3).



(a) Blurry frame (b) Colors shift (c) Lens contamination

Fig. 3. Example of difficult images in the test dataset: a significant frame blur caused by camera motion (a), a color components shift caused by the temporary signal failure (b) and an out-of-focus frame contains also contamination on the camera lens (c). Images taken from the CVC-12k [6].

From the RT-D approaches, the Xcept has the best overall performance, and it performs best on the CVC-12k dataset. The ResNe method reaches best performance for the Kvasir dataset, but is still far away from the GAN approach (MCC 0.262 versus 0.689).

The GF-D approach did not perform well on the CVC-12k dataset and could not make sense of the data. This is indicated by only negative MCC values which basically means no agreement. On the Kvasir dataset, it performed much better and could even outperform RT-D-VGG19. Overall, the RT-D approaches with VGG19 performed worse than all other approaches. The reason could be that the general hyper-parameters that we collected using optimization did not work well for the VGG19 architecture.

In order to compare our detection approaches to the state-of-the-art, we also evaluated one of the recent and promising object detection CNN called YOLOv2 [33]. The YOLOv2 model is able to detect objects within a frame and to provide an object's localization box and a probability value for the object detection. We trained YOLOv2 with the CVC-968 dataset using an appropriate conversion from ground truth masks to surrounding object boxes, as required by YOLOv2. The training was performed from scratch with the default model parameters. The trained YOLOv2 model showed relatively high performance with an MCC value of 0.450 and 0.296 for the Kvasir and CVC-12k sets, respectively, and was able to outperform all tested approaches except for the GAN-based solution. Nevertheless, the performance of the well-developed and already fine-tuned YOLOv2 model is significantly lower than our new GAN-based detection-via-localization approach.

TABLE V
RESULTS FOR THE FRAME-WISE POLYP DETECTION APPROACHES. WE USED THE CVC-12K AND KVASIR DATASET AS INDEPENDENT TEST SETS.

Test set	Run	Training set	PREC	SENS	SPEC	ACC	F1	MCC
Kvasir	GAN-356	CVC-356	0.715	0.751	0.940	0.909	0.732	0.677
	GAN-612	CVC-612	0.595	0.803	0.891	0.876	0.684	0.619
	GAN-968	CVC-968	0.736	0.746	0.946	0.913	0.741	0.689
	GF-D-356	CVC-356	0.171	0.109	0.894	0.763	0.133	0.004
	GF-D-612	CVC-612	0.270	0.318	0.828	0.743	0.292	0.137
	GF-D-968	CVC-968	0.225	0.859	0.409	0.484	0.357	0.208
	RT-D-Xcept-356	CVC-356	0.358	0.259	0.907	0.799	0.300	0.190
	RT-D-Xcept-612	CVC-612	0.383	0.326	0.895	0.800	0.352	0.236
	RT-D-Xcept-968	CVC-968	0.459	0.256	0.939	0.825	0.328	0.251
	RT-D-VGG19-356	CVC-356	0.181	0.333	0.777	0.720	0.235	0.087
	RT-D-VGG19-612	CVC-612	0.213	0.583	0.682	0.669	0.313	0.186
	RT-D-VGG19-968	CVC-968	0.231	0.320	0.842	0.774	0.268	0.142
	RT-D-ResNe-356	CVC-356	0.236	0.178	0.885	0.767	0.203	0.070
	RT-D-ResNe-612	CVC-612	0.321	0.507	0.785	0.739	0.393	0.247
	RT-D-ResNe-968	CVC-968	0.248	0.877	0.469	0.537	0.387	0.262
	YOLO-968	CVC-968	0.530	0.559	0.901	0.844	0.544	0.450
CVC-12k	GAN-356	CVC-356	0.967	0.624	0.888	0.667	0.758	0.378
	GAN-612	CVC-612	0.934	0.609	0.778	0.636	0.737	0.286
	GAN-968	CVC-968	0.906	0.912	0.510	0.847	0.909	0.428
	GF-D-356	CVC-356	0.829	0.909	0.030	0.767	0.867	-0.081
	GF-D-612	CVC-612	0.809	0.383	0.530	0.407	0.520	-0.064
	GF-D-968	CVC-968	0.835	0.854	0.125	0.737	0.845	-0.020
	RT-D-Xcept-356	CVC-356	0.913	0.624	0.693	0.636	0.742	0.236
	RT-D-Xcept-612	CVC-612	0.876	0.740	0.457	0.694	0.802	0.160
	RT-D-Xcept-968	CVC-968	0.899	0.690	0.600	0.676	0.781	0.224
	RT-D-VGG19-356	CVC-356	0.257	0.292	0.874	0.799	0.273	0.158
	RT-D-VGG19-612	CVC-612	0.266	0.489	0.799	0.759	0.344	0.228
	RT-D-VGG19-968	CVC-968	0.232	0.406	0.800	0.750	0.295	0.166
	RT-D-ResNe-356	CVC-356	0.723	0.003	0.999	0.871	0.006	0.044
	RT-D-ResNe-612	CVC-612	0.232	0.406	0.800	0.750	0.295	0.166
	RT-D-ResNe-968	CVC-968	0.870	0.303	0.766	0.378	0.450	0.057
	YOLO-968	CVC-968	0.932	0.641	0.757	0.660	0.759	0.296

V. CONCLUSIONS

In this paper, we have presented hand crafted and deep learning-based methods for automatic, pixel-, block- and frame-wise detection of polyps in videos from colonoscopies.

We evaluated the performance of our methods on different datasets. To achieve real-world comparability, we chose difficult datasets captured using different hardware equipment that were imbalanced in terms of positive, and negative examples and we also performed performance validation using different datasets for training and testing. Additionally, we tried to use as little amount of training data as possible. We showed that our newly proposed GAN based method outperforms handcrafted features and approaches based on well-known and working deep learning architectures. With our best working GAN-based approach, we reached detection specificity of 94% and accuracy of 90.9% with only 356 training and 6,000 test samples for the data captured by different equipment in different hospitals. The localization specificity and accuracy for the same training set are 98.4% and 94.6% respectively. Thus we can conclude that our approach works with a little amount of training data and, moreover, does not require negative examples for training, which is important to be able to use lesion imagery, already collected in hospitals. For future work, we plan to improve all methods presented in this paper with the main focus on the GAN-based approach, extend the experiments to other datasets and compare it to a broader range of approaches including a time-series-based analysis using for example long short-term memory.

REFERENCES

[1] M. F. Kaminski, J. Regula, E. Kraszewska, M. Polkowski, U. Wojciechowska, J. Didkowska, M. Zwierko, M. Rupinski, M. P. Nowacki, and E. Butruk, "Quality indicators for colonoscopy and the risk of interval cancer," *New England Journal of Medicine*, vol. 362, no. 19, pp. 1795–1803, 2010.

[2] World Health Organization - International Agency for Research on Cancer, "Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012," 2012.

[3] H. Brenner, M. Kloor, and C. P. Pox, "Colorectal cancer," *Lancet*, vol. 383, no. 9927, pp. 1490–502, 2014.

[4] J. C. van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. van Deventer, and E. Dekker, "Polyp miss rate determined by tandem colonoscopy: a systematic review," *The American journal of gastroenterology*, vol. 101, no. 2, pp. 343–350, 2006.

[5] The New York Times, "The $2.7 Trillion Medical Bill," http://goo.gl/CuFyFJ, [last visited, Nov. 29, 2015].

[6] J. Bernal and H. Aymeric, "Miccai endoscopic vision challenge polyp detection and segmentation," https://endovissub2017-giana.grand-challenge.org/home/, accessed: 2017-12-11.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[9] M. Riegler, C. Griwodz, C. Spampinato, T. de Lange, S. L. Eskeland, K. Pogorelov, W. Tavanapong, P. T. Schmidt, C. Gurrin, D. Johansen, H. Johansen, and P. Halvorsen, "Multimedia and medicine: Teammates for better disease detection and survival," in *Proc. of ACM MM*, 2016, pp. 968–977.

[10] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of MMSYS*, june 2017, pp. 164–169.

[11] A. Mamonov, I. Figueiredo, P. Figueiredo, and Y.-H. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, July 2014.

[12] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2014.

[13] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. De Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *Computer methods and programs in biomedicine*, vol. 120, no. 3, pp. 164–179, 2015.

[14] Y. Yuan, D. Li, and M. Q.-H. Meng, "Automatic polyp detection via a novel unified bottom-up and top-down saliency approach," *IEEE Journal of Biomedical and Health Informatics*, 2017.

[15] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[16] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[17] Y. Shin and I. Balasingham, "Comparison of hand-craft feature based svm and cnn based deep learning framework for automatic polyp classification," in *Proc. of EMBC*, 2017, pp. 3277–3280.

[18] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proc. of IEEE ISBI*, 2015, pp. 79–83.

[19] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 65–75, 2017.

[20] K. Pogorelov, M. Riegler, S. L. Eskeland, T. de Lange, D. Johansen, C. Griwodz, P. T. Schmidt, and P. Halvorsen, "Efficient disease detection in gastrointestinal videos–global features versus neural networks," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22 493–22 525, 2017.

[21] M. Riegler, K. Pogorelov, S. L. Eskeland, P. T. Schmidt, Z. Albisser, D. Johansen, C. Griwodz, P. Halvorsen, and T. D. Lange, "From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3, p. 26, 2017.

[22] V. Baptista, N. Marya, A. Singh, A. Rupawala, B. Gondal, and D. Cave, "Continuing challenges in the diagnosis and management of obscure gastrointestinal bleeding," *World journal of gastrointestinal pathophysiology*, vol. 5, no. 4, p. 523, 2014.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014.

[24] J. Son, S. J. Park, and K.-H. Jung, "Retinal vessel segmentation in fundoscopic images with generative adversarial networks," *arXiv preprint arXiv:1706.09318*, 2017.

[25] M. Lux, M. Riegler, P. Halvorsen, K. Pogorelov, and N. Anagnostopoulos, "Lire: open source visual information retrieval," in *Proc. of ACM MMSys*, 2016.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[28] K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Nerthus: A bowel preparation quality video dataset," in *Proc. of MMSYS*, june 2017, pp. 170–174.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint arXiv:1610.02357*, 2016.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE CVPR*, 2016, pp. 770–778.

[32] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[33] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.